Unit 5: Hypothesis Testing

The Practice of Statistics, 4th edition – For AP*

STARNES, YATES, MOORE

Unit 5: Hypothesis Testing

- 9.1 Significance Tests: The Basics
- 9.2 Tests about a Population Proportion
- **9.3 Tests about a Population Mean**
- 9.1 & 9.2 Errors and the Power of a Test

Section 9.3
Tests About a Population Mean

Learning Objectives

After this section, you should be able to…

- ✓ CHECK conditions for carrying out a test about a population mean.
- ✓ CONDUCT a one-sample *t* test about a population mean.
- ✓ CONSTRUCT a confidence interval to draw a conclusion for a two-sided test about a population mean.

Introduction

Confidence intervals and significance tests for a population proportion p are based on z-values from the standard Normal distribution.

Inference about a population mean μ uses a t distribution with $n - 1$ degrees of freedom, except in the rare case when the population standard deviation σ is known.

Carrying Out a Significance Test for μ

In an earlier example, a company claimed to have developed a new AAA battery that lasts longer than its regular AAA batteries. Based on years of experience, the company knows that its regular AAA batteries last for 30 hours of continuous use, on average. An SRS of 15 new batteries lasted an average of 33.9 hours with a standard deviation of 9.8 hours. Do these data give *convincing evidence* that the new batteries last longer on average?

To find out, we must perform a significance test of

$$H_0: \mu = 30 \text{ hours}$$
$$H_a: \mu > 30 \text{ hours}$$

where μ = the true mean lifetime of the new deluxe AAA batteries.

Check Conditions:

Three conditions should be met before we perform inference for an unknown population mean: Random, Normal, and Independent. The Normal condition for means is

Population distribution is Normal or sample size is large ($n \geq 30$)

We often don't know whether the population distribution is Normal. But if the sample size is large ($n \geq 30$), we can safely carry out a significance test (due to the central limit theorem). If the sample size is small, we should examine the sample data for any obvious departures from Normality, such as skewness and outliers.

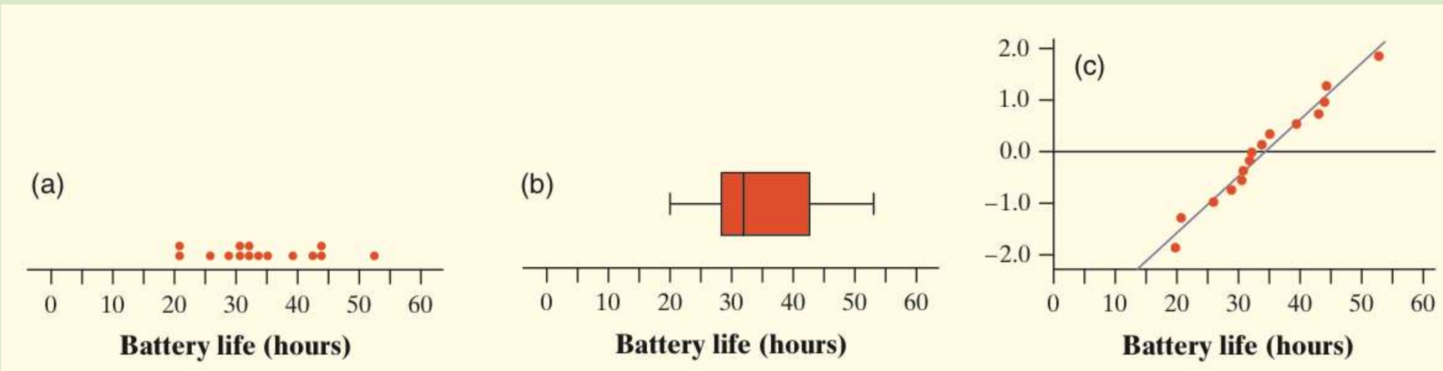
Carrying Out a Significance Test for μ

Check Conditions:

Three conditions should be met before we perform inference for an unknown population mean: Random, Normal, and Independent.

✓ ***Random*** The company tests an SRS of 15 new AAA batteries.

✓***Normal*** We don't know if the population distribution of battery lifetimes for the company's new AAA batteries is Normal. With such a small sample size (n = 15), we need to inspect the data for any departures from Normality.



The dotplot and boxplot show slight right-skewness but no outliers. The Normal probability plot is close to linear. We should be safe performing a test about the population mean lifetime μ.

✓***Independent*** Since the batteries are being sampled without replacement, we need to check the *10% condition*: there must be at least 10(15) = 150 new AAA batteries. This seems reasonable to believe.

Carrying Out a Significance Test

Calculations: Test statistic and P-value

When performing a significance test, we do calculations assuming that the null hypothesis H_0 is true. The test statistic measures how far the sample result diverges from the parameter value specified by H_0, in standardized units. As before,

$$\text{test statistic} = \frac{\text{statistic - parameter}}{\text{standard deviation of statistic}}$$

For a test of H_0: $\mu = \mu_0$, our statistic is the sample mean. Its standard deviation is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

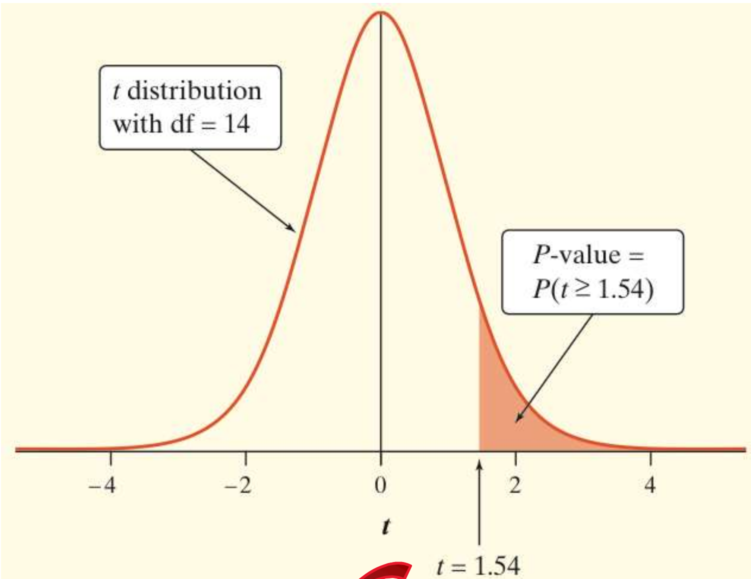
Because the population standard deviation σ is usually unknown, we use the sample standard deviation s_x in its place. The resulting test statistic has the standard error of the sample mean in the denominator

$$t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$$

When the Normal condition is met, this statistic has a *t* distribution with *n* - 1 degrees of freedom.

Carrying Out a Hypothesis Test

The battery company wants to test H_0: $\mu = 30$ versus H_a: $\mu > 30$ based on an SRS of 15 new AAA batteries with mean lifetime and standard deviation $\bar{x} = 33.9$ hours and $s_x = 9.8$ hours.



$$\text{test statistic} = \frac{\text{statistic - parameter}}{\text{standard deviation of statistic}}$$

$$t = \frac{\bar{x} - \mu_0}{s_x / \sqrt{n}} = \frac{33.9 - 30}{9.8/\sqrt{15}} = 1.54$$

The *P*-value is the probability of getting a result this large or larger in the direction indicated by H_a, that is, $P(t \geq 1.54)$.

Upper-tail probability *p*

df	**.10**	**.05**	**.025**
13	1.350	1.771	2.160
14	1.345	1.761	2.145
15	1.341	1.753	3.131
	80%	90%	95%

Confidence level *C*

✓ Go to the *df* = 14 row.

✓ Since the *t* statistic falls between the values 1.345 and 1.761, the "Upper-tail probability *p*" is between 0.10 and 0.05.

✓ The *P*-value for this test is between 0.05 and 0.10.

Because the *P*-value exceeds our default α = 0.05 significance level, we can't conclude that the company's new AAA batteries last longer than 30 hours, on average.

Using Table B Wisely

- Table B gives a range of possible *P*-values for a significance. We can still draw a conclusion from the test in much the same way as if we had a single probability by comparing the range of possible *P*-values to our desired significance level.

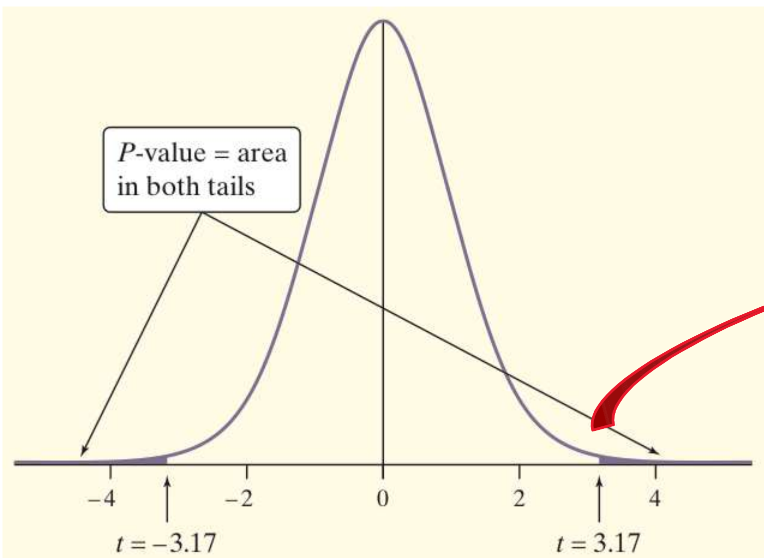
- Table B has other limitations for finding P-values. It includes probabilities only for *t* distributions with degrees of freedom from 1 to 30 and then skips to *df* = 40, 50, 60, 80, 100, and 1000. (The bottom row gives probabilities for *df* = ∞, which corresponds to the standard Normal curve.) *Note: If the df you need isn't provided in Table B, use the next lower df that is available.*

- Table B shows probabilities only for positive values of *t*. To find a *P*-value for a negative value of *t*, we use the symmetry of the *t* distributions.

Using Table B Wisely

Suppose you were performing a test of H_0: $\mu = 5$ versus H_a: $\mu \neq 5$ based on a sample size of $n = 37$ and obtained $t = -3.17$. Since this is a two-sided test, you are interested in the probability of getting a value of t less than -3.17 or greater than 3.17.

Due to the symmetric shape of the density curve, $P(t \leq -3.17) = P(t \geq 3.17)$. Since Table B shows only positive t-values, we must focus on $t = 3.17$.



Upper-tail probability *p*

df	**.005**	**.0025**	**.001**
29	2.756	3.038	3.396
30	2.750	3.030	3.385
40	2.704	2.971	3.307
	99%	99.5%	99.8%

Confidence level *C*

Since $df = 37 - 1 = 36$ is not available on the table, move across the $df = 30$ row and notice that $t = 3.17$ falls between 3.030 and 3.385.
The corresponding "Upper-tail probability p" is between 0.0025 and 0.001. For this two-sided test, the corresponding P-value would be between 2(0.001) = 0.002 and 2(0.0025) = 0.005.

Example: Healthy Streams

The level of dissolved oxygen (DO) in a stream or river is an important indicator of the water's ability to support aquatic life. A researcher measures the DO level at 15 randomly chosen locations along a stream. Here are the results in milligrams per liter:

4.53	5.04	3.29	5.23	4.13	5.50	4.83	4.40
5.42	6.38	4.01	4.66	2.87	5.73	5.55	

A dissolved oxygen level below 5 mg/l puts aquatic life at risk.

State: We want to perform a test at the $\alpha = 0.05$ significance level of

$$H_0: \mu = 5$$
$$H_a: \mu < 5$$

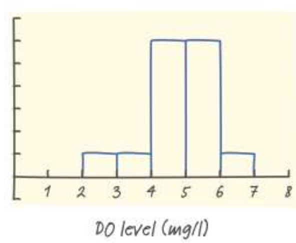
where μ is the actual mean dissolved oxygen level in this stream.

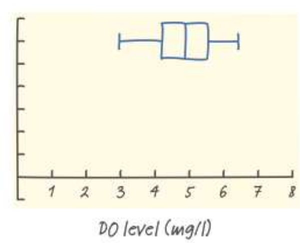
Plan: If conditions are met, we should do a one-sample t test for μ.

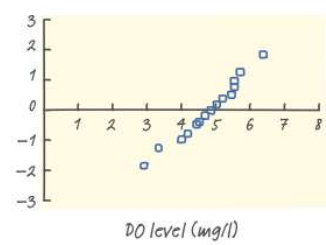
✓*Random* The researcher measured the DO level at 15 randomly chosen locations.

✓*Normal* We don't know whether the population distribution of DO levels at all points along the stream is Normal. With such a small sample size ($n = 15$), we need to look at the data to see if it's safe to use t procedures.

The histogram looks roughly symmetric; the boxplot shows no outliers; and the Normal probability plot is fairly linear. With no outliers or strong skewness, the t procedures should be pretty accurate even if the population distribution isn't Normal.







Example: Healthy Streams

The level of dissolved oxygen (DO) in a stream or river is an important indicator of the water's ability to support aquatic life. A researcher measures the DO level at 15 randomly chosen locations along a stream. Here are the results in milligrams per liter:

4.53	5.04	3.29	5.23	4.13	5.50	4.83	4.40
5.42	6.38	4.01	4.66	2.87	5.73	5.55	

A dissolved oxygen level below 5 mg/l puts aquatic life at risk.

State: We want to perform a test at the $\alpha = 0.05$ significance level of

$$H_0: \mu = 5$$
$$H_a: \mu < 5$$

where μ is the actual mean dissolved oxygen level in this stream.

Plan: If conditions are met, we should do a one-sample t test for μ.

✓*Random* The researcher measured the DO level at 15 randomly chosen locations.

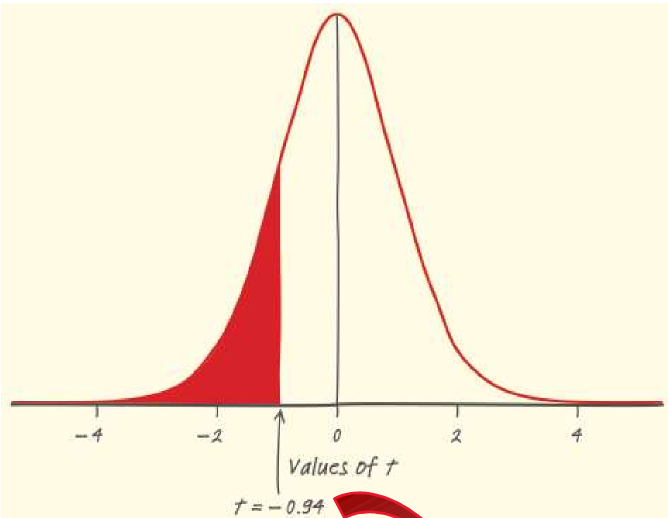
✓*Normal* We don't know whether the population distribution of DO levels at all points along the stream is Normal. With such a small sample size ($n = 15$), we need to look at the data to see if it's safe to use t procedures.

The histogram looks roughly symmetric; the boxplot shows no outliers; and the Normal probability plot is fairly linear. With no outliers or strong skewness, the t procedures should be pretty accurate even if the population distribution isn't Normal.

✓*Independent* There is an infinite number of possible locations along the stream, so it isn't necessary to check the 10% condition. We do need to assume that individual measurements are independent.

Example: Healthy Streams

Do: The sample mean and standard deviation are $\bar{x} = 4.771$ and $s_x = 0.9396$



Test statistic $t = \dfrac{\bar{x} - \mu_0}{s_x / \sqrt{n}} = \dfrac{4.771 - 5}{0.9396 / \sqrt{15}} = -0.94$

***P*-value** The *P*-value is the area to the left of *t* = -0.94 under the *t* distribution curve with df = 15 – 1 = 14.

Upper-tail probability *p*

df	.25	.20	.15
13	.694	.870	1.079
14	.692	.868	1.076
15	.691	.866	1.074
	50%	60%	70%

Confidence level *C*

Conclude: The *P*-value, is between 0.15 and 0.20. Since this is greater than our α = 0.05 significance level, we fail to reject H_0. We don't have enough evidence to conclude that the mean DO level in the stream is less than 5 mg/l.

Since we decided not to reject H_0, we could have made a Type II error (failing to reject H_0 when H_0 is false). If we did, then the mean dissolved oxygen level μ in the stream is actually less than 5 mg/l, but we didn't detect that with our significance test.

Two-Sided Tests

At the Hawaii Pineapple Company, managers are interested in the sizes of the pineapples grown in the company's fields. Last year, the mean weight of the pineapples harvested from one large field was 31 ounces. A new irrigation system was installed in this field after the growing season. Managers wonder whether this change will affect the mean weight of future pineapples grown in the field. To find out, they select and weigh a random sample of 50 pineapples from this year's crop. The Minitab output below summarizes the data. Determine whether there are any outliers.

Descriptive Statistics: Weight (oz)

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Weight (oz)	50	31.935	0.339	2.394	26.491	29.990	31.739	34.115	35.547

✓ $IQR = Q_3 - Q_1 = 34.115 - 29.990 = 4.125$

✓ Any data value greater than $Q_3 + 1.5(IQR)$ or less than $Q_1 - 1.5(IQR)$ is considered an outlier.

$$Q_3 + 1.5(IQR) = 34.115 + 1.5(4.125) = 40.3025$$
$$Q_1 - 1.5(IQR) = 29.990 - 1.5(4.125) = 23.0825$$

✓ Since the maximum value 35.547 is less than 40.3025 and the minimum value 26.491 is greater than 23.0825, there are no outliers.

Two-Sided Tests

State: We want to test the hypotheses

$$H_0: \mu = 31$$
$$H_a: \mu \neq 31$$

where μ = the mean weight (in ounces) of all pineapples grown in the field this year. Since no significance level is given, we'll use $\alpha = 0.05$.

Plan: If conditions are met, we should do a one-sample *t* test for μ.

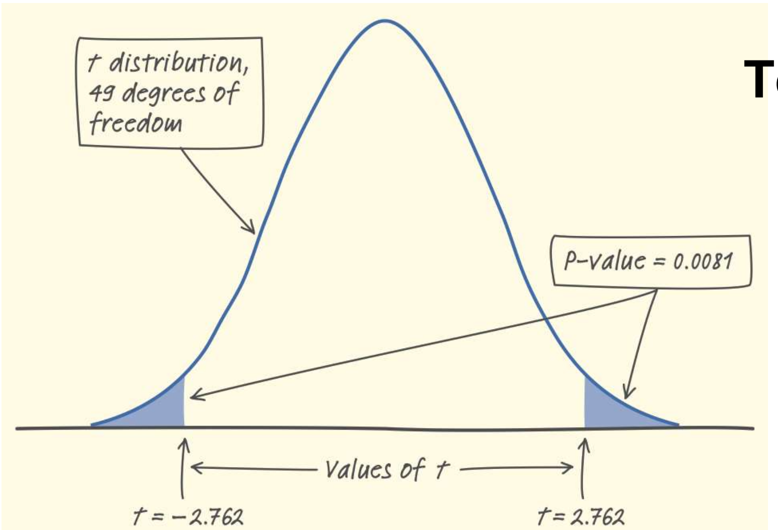
✓*Random* The data came from a random sample of 50 pineapples from this year's crop.

✓*Normal* We don't know whether the population distribution of pineapple weights this year is Normally distributed. But $n = 50 \geq 30$, so the large sample size (and the fact that there are no outliers) makes it OK to use *t* procedures.

✓*Independent* There need to be at least 10(50) = 500 pineapples in the field because managers are sampling without replacement (*10% condition*). We would expect many more than 500 pineapples in a "large field."

Two-Sided Tests

Do: The sample mean and standard deviation are $\bar{x} = 31.935$ and $s_x = 2.394$



Test statistic $t = \dfrac{\bar{x} - \mu_0}{s_x / \sqrt{n}} = \dfrac{31.935 - 31}{2.394 / \sqrt{50}} = 2.762$

***P*-value** The *P*-value for this two-sided test is the area under the *t* distribution curve with 50 - 1 = 49 degrees of freedom. Since Table B does not have an entry for df = 49, we use the more conservative df = 40. The upper tail probability is between 0.005 and 0.0025 so the desired *P*-value is between 0.01 and 0.005.

Upper-tail probability *p*

df	**.005**	**.0025**	**.001**
30	2.750	3.030	3.385
40	2.704	2.971	3.307
50	2.678	2.937	3.261
	99%	99.5%	99.8%

Confidence level *C*

Conclude: Since the *P*-value is between 0.005 and 0.01, it is less than our *α* = 0.05 significance level, so we have enough evidence to reject H_0 and conclude that the mean weight of the pineapples in this year's crop is not 31 ounces.

Confidence Intervals Give More Information

Minitab output for a significance test and confidence interval based on the pineapple data is shown below. The test statistic and *P*-value match what we got earlier (up to rounding).

One-Sample T: Weight (oz)

Test of mu = 31 vs not = 31

Variable	N	Mean	StDev	SE Mean	95% CI	T	P
Weight (oz)	50	31.935	2.394	0.339	(31.255, 32.616)	2.76	0.008

The 95% confidence interval for the mean weight of all the pineapples grown in the field this year is 31.255 to 32.616 ounces. We are 95% confident that this interval captures the true mean weight μ of this year's pineapple crop.

As with proportions, there is a link between a two-sided test at significance level α and a 100(1 – α)% confidence interval for a population mean μ.

For the pineapples, the two-sided test at $\alpha = 0.05$ rejects H_0: $\mu = 31$ in favor of H_a: $\mu \neq 31$. The corresponding 95% confidence interval does not include 31 as a plausible value of the parameter μ. In other words, the test and interval lead to the same conclusion about H_0. But the confidence interval provides much more information: *a set of plausible values for the population mean*.

Confidence Intervals and Two-Sided Tests

The connection between two-sided tests and confidence intervals is even stronger for means than it was for proportions. That's because both inference methods for means use the standard error of the sample mean in the calculations.

Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$ Confidence interval: $\bar{x} \pm t^* \dfrac{s_x}{\sqrt{n}}$

✓ A two-sided test at significance level α (say, $\alpha = 0.05$) and a $100(1 - \alpha)\%$ confidence interval (a 95% confidence interval if $\alpha = 0.05$) give similar information about the population parameter.

✓ When the two-sided significance test at level α rejects H_0: $\mu = \mu_0$, the $100(1 - \alpha)\%$ confidence interval for μ will not contain the hypothesized value μ_0 .

✓ When the two-sided significance test at level α fails to reject the null hypothesis, the confidence interval for μ will contain μ_0 .